DOCUMENT RESUME

ED 377 211 TM 022 376

AUTHOR Campbell, Todd C.
TITLE Factors That Do and Do Not Affect "r" and Its Generalizations.
PUB DATE Apr 94
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Behavioral Science Research; *Correlation; *Generalization; Scores
IDENTIFIERS Additive Models; Multiplicative Strategies; *Pearson Product Moment Correlation

ABSTRACT

Correlation is one of the most widely used analytic procedures in the behavioral sciences. The bivariate correlation is implicit in all classical analyses ranging from t-tests to canonical correlation analysis. The most common correlation coefficient used in statistics is the Pearson product-moment coefficient of correlation, which is represented by the symbol "r." The present paper discusses factors that do and do not affect "r" and its generalizations, including additive and multiplicative constants. Seven tables and 11 figures are included. (Contains 20 references.) (Author)

Factors That Do and Do Not Affect r and Its Generalizations

Todd C. Campbell

Texas A & M University 77843-4225

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April 6.1994.

Abstract

Correlation is one of the most widely used analytic procedures in the behavioral sciences. The bivariate correlation is implicit in all classical analyses ranging from t-tests to canonical correlation analysis. The most common correlation coefficient used in statistics is the Pearson product-moment coefficient of correlation, which is represented by the symbol r. The present paper discusses the factors that do and do not affect r and its generalizations, including additive and multiplicative constants.

Correlation is one of the most widely used analytic procedures in the behavioral sciences in both published research (Edington, 1964, 1974; Elmore & Wolke, 1988; Goodwin & Goodwin, 1985; Willson, 1980) and dissertation research (cf. Lagaccia, 1991; Wick & Dirkes, 1973). The bivariate correlation coefficient is implicit in all classical parametric analyses ranging from t-tests to canonical correlation analysis (Knapp, 1978; Thompson, 1991). Therefore, it is essential that students of social science research be cognizant of the factors that do and do not affect correlations and the generalizations made from the interpretations of correlations.

The most common correlation coefficient used in statistics is the Pearson product-moment coefficient of correlation, which is represented by the symbol, r, for a sample (rho, for the population parameter). In fact, r is employed so often that unless another coefficient is specified, the term "correlation" is assumed to mean the Pearson product-moment coefficient of correlation (Nunnally, 1967). Glass and Hopkins (1984) define the correlation coefficient, r, as "a statistical summary of the degree and direction of relationship or association between two variables" (p. 79).

All correlations measure the strength of relationship between two sets of scores. The scores must be paired observations (Hinkle, Wiersma & Jurs, 1994). For example, we cannot determine a correlation coefficient for socioeconomic status (SES) for one group of subjects with the level of education for another completely different group of subjects. SES and the level of education scores must both be available for each subject in the analysis.

One formula for calculating the Pearson product moment coefficient of correlation for a sample, r, is:

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \, \Sigma y^2}}$$

Because the mean and standard deviation are used to compute r, we must use interval or ratio data in calculating r (Hinkle, Wiersma & Jurs, 1994). Nevertheless, there are special cases of the Pearson product-moment correlation, r, that can be utilized with lower levels of measurement, i.e., nominal and ordinal data. The phi, Spearman's rho, biserial, and point biserial coefficients are

common variations of r that are applicable to nominal and ordinal data (McCallister, 1991). However, the present paper will focus on r and the factors that do and do not affect r .

Linearity

In determining the strength of a correlation there is an underlying assumption of a linear relationship between the variables (Nunnally, 1967). The relationship between the data can be illustrated graphically on a scatterplot with the horizontal axis labeled X and the vertical axis labeled Y, by convention. The scatterplot provides a visual aid to determine if the relationship between the variables is linear.

The "best fitting" regression line through a scatterplot must be a straight line. If the data create a curved pattern in the scattergram, then r will underestimate the degree of relationship between the two variables (Nunnally, 1967). For example, the relationship between physical endurance and age is a curvilinear relationship. In the earliest stages of life there is little endurance, as compared to adolescence and young adulthood when physical endurance tends to increase, but as one enters middle age and late adulthood physical endurance tends to decrease. This relationship would be plotted on a scattergram as curvilinear, i.e., an inverted u-shaped relationship. The degree to which such a relationship is underestimated is commensurate to the degree of curvilinearity within the relationship (Murthy, 1993).

Insert Table 1 and Figure 1 about here.

Welkowitz, Ewen and Cohen (1982) make the point that, strictly speaking, there need be no assumption of linearity " ... if r is considered a measure of the degree of linear relationship, it remains such a measure whether or not the best-fitting function is linear " (p. 185). The authors then qualify this statement, noting that, "Ordinarily, however, one would not be interested in the best linear fit when it is known that the relationship is not linear" (p. 185).

The Range of Values For r and Interpretation of the Values

The correlation coefficient, r , can range in value from -1.00 to +1.00 inclusive, with the

extreme scores representing a perfect linear relationship between the variables. A positive score means that as individuals score above the mean on one variable they tend to score above the mean on a corresponding variable or as individuals score below the mean on one variable they tend to score below the mean on the corresponding variable. A score of +1.00 represents a perfect positive linear correlation between the observed variables.

Insert Table 2 and Figure 2 about here.

A negative score means that as individuals score above the mean on one variable they tend to score below the mean on a corresponding variable or, conversely, as individuals score below the mean on one variable they tend to score above the mean on the corresponding variable. A value of -1.00 represents a perfect negative linear correlation between the observed variables.

Insert Table 3 and Figure 3 about here.

If $r = 0.00$, this represents the absence of any linear relationship between the variables, though there may be a strong curvilinear relationship (Murthy, 1993). A useless predictor variable will also produce an $r = 0.00$. For instance, the r between waist size and IQ might be zero. Logically, waist size and IQ are unrelated, so an r of zero between these two variables might be expected. The linear regression formula also will not predict any linear relationship when there is a useless variable involved in the calculation. Welkowitz, Ewen and Cohen (1982) explain,

> If there is no good information on which to base a prediction, the same estimate -the mean of the criterion--is made for everyone. When r between the variables is equal to 0.00, the linear regression formula becomes (p.191):

$$b_{yx} = .00\,(\sigma_y / \sigma_x) = 0$$

$$a_{yx} = \overline{Y} - (0.00)\,\overline{X} = \overline{Y}$$

$$Y' = b_{yx}X + a_{yx} = (0)\,X + \overline{Y} = \overline{Y}$$

Insert Table 4 Figure 4 about here.

The direction (positive or negative) of the correlation is dependent upon how the variables are framed. Nunnally (1975) gives the example that correlating the number of errors on a spelling test with the number of correct answers on a math test may yield a correlation of (-.72). However, if the test was reframed as the correlation between the number of correct answers on a spelling test and the number of correct answers on a math test the correlation may "turn" positive and yield an r = (+.72) (p.144).

The correlation coefficient is represented on an ordinal scale. It is a relative measure of the relationship, not an absolute measure. That is, a score of .80 is not twice as strong as a score of .40 (Hinkle, Wiersma & Jurs, 1994).

The Coefficient of Determination and r

A common misconception held by novice researchers is that r represents the amount of common variance of two variables. In order to find the common variance we need to compute the coefficient of determination, r squared, which represents the proportion of the total variance in Y that can be associated with the variance in X. Spatz and Johnston (1984) provide the example of a novice researcher concluding that r = .70 is a very high correlation, but in actuality only about half the variance is held in common or explained. The researcher would need an r = .84 to predict 70% of the variance (.84 squared = .7056). A correlation coefficient equal to .84 would certainly be a very high correlation for social science research. In contrast to r, the coefficient of determination, r squared, does allow us to compare different coefficients by proportion because coefficients of determination are on a linear scale.

We can represent, with a Venn diagram, the variance of each variable and the variance that is held in common. The size of each circle in the Venn diagram represents the variance of a variable. The overlapping areas of the circles represent the common or explained variance, that is, the coefficient of determination, r squared.

Insert Figure 5 about here.

The Effects of Additive and Multiplicative Constants on r

It is often desirable to transform scores so that the sets of scores of interest have equal means and standard deviations. This transformation is obtained by adding or subtracting a constant from every score and/or multiplying or dividing every score by a constant. This transformation allows for accurate comparison of scores. The transformation of scores into standardized Z scores, which have a mean of 0 and a standard deviation of 1, not only allows for the accurate comparison of scores, but also provides a simple method of determining whether or not a score is above or below the mean and how many standard deviations above or below the mean the score falls. Z scores also allow us to infer, from the normal curve, the percentage of the population that scores above or below that particular score, if the scores are normally distributed.

It is important to understand how the basic building blocks of statistical analysis, the measures of central tendency and variation, are affected by these progressive abstractions of the data. Because the correlation coefficient formula utilizes the mean and the standard deviation, it is also important to understand how r is or is not affected by these transformations.
It can be shown that:

Additive Constants.

1. Adding a constant to a set of scores raises the mean by the value of the constant.
2. Subtracting a constant from a set of scores decreases the mean by the value of the constant.
3. Adding or subtracting a constant value from each of a set of

scores does not affect the variance or standard deviation of the set of scores. This can be grasped intuitively. Because each score is changed by the same amount the set of scores is "moved" as a whole and the "spreadoutness" is not affected.

4. **The correlation coefficient is not changed** by adding or subtracting a constant value from each score

Multiplicative Constants.

1. Multiplying each score by a constant value results in a new mean for a set of scores which is equal to the old mean multiplied by the value of the constant.
2. Dividing each score by a constant value results in a new mean for a set of scores which is equal to the old mean divided by the value of the constant.
3. Multiplying each score by a constant value results in a new standard deviation for a set of scores which is equal to the old standard deviation multiplied by the value of the constant.
4. Dividing each score by a constant value results in a new standard deviation for a set of scores which is equal to the old standard deviation divided by the value of the constant.
5. **The correlation coefficient is not changed** by multiplying or dividing each score by a constant value.

Insert Figure 6 about here.

Correlation and Causality

The correlation coefficient does not provide evidence of causality. The correlation coefficient is only a measure of the linear relationship between the two variables. A classic

example of this concept is the high positive correlation between the number of storks sighted and the number of births in a European city. One might jump to the conclusion that storks cause the birth of children. Then one might propose that the elimination of the stork population would result in effective birth control for humans. A responsible researcher would not jump to such conclusions regarding causality and would search for other factors that may explain the relationship between the variables. One explanation may be that as the population increased the number of homes increased and in turn the number of chimneys in the community increased, which just happened to be the favorite nesting places for storks. This explains the positive correlation between the number of storks and the number of births.

Spurious Correlations

Variables may seem to co-vary in relation to each other, but there may be a hidden variable or lurking variable. Johnson and Bhattacharyya (1985) cite the example of recording the number of homicides (x) in a city and the number of religious meetings (y) in the same city. The data will probably produce a high positive correlation between the variables, but it is actually the fluctuation of a third variable, a lurking variable, i.e., the city's population, that causes x and y to vary in the same direction. It may be that the variables actually have a very weak correlation or even a strong negative correlation. The deceptive correlation that results is called a spurious correlation (Johnson & Bhattacharyya, 1985). The problem of spurious correlation is a logical and not a statistical problem.

If the sample contains subgroups that have different means or standard deviations a spurious correlation may result (Kirk, 1984). The r may appear substantially greater or lesser if two subgroups with different means are combined than if the subgroups were investigated separately. Kirk (1984) cites the example: If we are interested in the correlation between anxiety level and school achievement, two subgroups may emerge, i.e., low income and middle income students. If the mean score for the subjects from the middle income families is greater than the

mean score for the subjects in the lower income families with respect to both anxiety (X) and achievement (Y) then a spuriously high correlation will result. If the mean scores of the two subgroups differ on only one of the variables, for example, anxiety (X) then r for the combined sample will be smaller than if r was computed for the subgroups separately.

Insert Figure 7 about here.

If the sample contains subgroups that have different standard deviations, but do not differ in their mean scores in respect to one or both of the variables then a spurious correlation can result (Kirk, 1984). This is illustrated in Figure 8.

Insert Figure 8 about here.

Another example of when the researcher's judgment is needed is when the data break into two clusters on the scatterplot. This may mean that the sample is actually two samples from different populations and r should not be used to interpret the data. It is necessary to determine the underlying cause of such splits in the data (Johnson & Bhattacharyya, 1985).

Insert Table 5 and Figure 9 about here.

Dissimilar Distributions and the Affects on r

It is also assumed that the observed variables share the same shape in their distributions. The more dissimilar the shapes of the distributions, the more restricted is the value of r (Dolenz-Walsh, 1992). If the shapes of the distributions are dissimilar, it is impossible to obtain a perfect r, $r = \pm 1.00$ (Murthy, 1993; Nunnally, 1975). Nunnally (1975) cites an extreme example of correlating a continuous variable with a dichotomous variable. The dichotomous variable, by definition, is split into two categories and the continuous variable, if normally distributed, will be spread over the continuum. To obtain a perfect correlation all of the continuous variable scores

will have to fall on the two points of the dichotomous variable scores. Logically this cannot happen. The actual relationship between a dichotomous variable and a continuous variable can be shown graphically (Nunnally, 1975, p. 154).

Insert Table 6 and Figure 10 about here.

Nunnally (1975) states that dissimilar shapes do not greatly affect correlations of .50 or less and correlations of .30 or less are hardly affected by even drastic differences in the shapes of the distributions. He goes on to state that seemingly small correlations are very common in psychology noting that. ". . . the average of all correlations reported in the literature is probably less than 0.40" (p. 155), so unless a drastic difference is detected by the naked eye the researcher can move ahead confidently.

It is assumed that the distributions are homoscedastic, that is, the variation is approximately equal along the best fitting regression line in a scattergram (Nunnally, 1975; Walsh, 1992). If the distribution is heteroscadastic, that is, the variation differs along the best fitting regression line, then more precise predictions can be made at the points of lesser variation and in turn, less precise predictions can be made at points of greater variation.

Insert Table 7 and Figure 11 about here.

A homogeneous group results in a restricted range for a variable (Hinkle, Wiersma & Jurs, 1994). As homogeneity increases variability decreases. If either variable is truncated in range then the size of r will be reduced (Kirk, 1984). If the group is absolutely homogeneous then the standard deviation on that variable for the group is equal to zero, which renders the formula for r meaningless because we are not allowed to divide by zero. This is also a logical issue because if a group is the same on all counts as regards one variable then there is no point in trying to determine if there is a correlation between the two sets of scores.

Kirk (1984) states that the problem of restricted range is a common problem in educational and behavioral research. This is due to the fact that much of the research conducted in these areas utilizes college students as their subjects of investigation. Since college students tend to be a relatively homogeneous population, particularly in the areas of intelligence and age, this group tends to be restricted in range thus often reducing the size of r.

Other Factors That Do or Do Not Affect r

The size of the sample does not affect the size of r (except when n = 2), but does affect the accuracy of r (Hinkle, Wiersma & Jurs, 1994). Measurement error can lead to the attenuation of r (Busby & Thompson, 1990). It is important to assess the reliability coefficients for the scores in hand because the reliability coefficients for the variables being studied establishes a ceiling for the value of r. The value of r can never exceed the square root of the product of the two reliability coefficients for the scores of the variables being studied (Busby & Thompson, 1990).

Conclusion

There is a wide array of factors that do and do not affect r and its generalizations. Many of these factors have been discussed in the present paper. Considering the integral part that the Pearson product-moment coefficient of correlation (r) has in behavioral science research, it is important that the student of this type of research be aware of the factors that do and do not affect r. Not only is the researcher's cognizance of these factors required to evaluate r in a meaningful way, but the researcher's judgment must also be called into play.

References

Busby, D., & Thompson, B. (1990, January). Factors attenuating Pearson's r: A review of basics and some corrections. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

Dolenz-Walsh, B. (1992, January) Factors that attenuate the correlation coefficient and its analogs. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 347 173)

Edgington, E.S. (1964). A tabulation of inferential statistics used in psychology journals. American Psychologist, 19, 202-203.

Edgington, E.S. (1974). A new tabulation of statistical procedures in APA journals. American Psychologist, 29 (1), 25-26.

Elmore, P.B., & Woehlke, P.L. (1988). Statistical methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987. Educational Researcher, 17 (9), 19-20.

Glass, G.V., & Hopkins, K.D. (1984). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall Inc.

Goodwin, L.D., & Goodwin, W.L. (1985). Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 14 (2), 5-11.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (1994). Applied statistics for the behavioral sciences. Boston, MA: Houghton Mifflin Company.

Johnson, R., & Bhattacharyya, G. (1985). Statistics: Principles and methods. New York: John Wiley & Sons.

Kirk, R. E. (1984). Elementary Statistics (2nd ed.). Belmont, CA: Brooks/Cole Publishing.

Knapp, T.R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Lagaccia, S.S. (1991). Methodology choices in a cohort of education dissertations. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 149-158). Greenwhich, CT: JAI Press.

Murthy, K. (1993, November). What makes r positive or negative?: An exploration of factors that affect r with an emphasis on insight and understanding. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA.

Nunnally, J. C. (1975). Introduction to statistics for psychology and education. New York: McGraw-Hill Book Company.

Nunnally, J.C. (1967). Psychometric theory. New York: McGraw-Hill Book Company.

Spatz, C., & Johnston, J.O. (1984). Basic statistics: Tales of distributions. Belmont, CA: Brooks/Cole Publishing Company.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24 (2), 80-95.

Welkowitz, J., Ewen, R.B., & Cohen, J. (1982). Introductory statistics for the behavioral sciences (3rd ed.). New York: Academic Press Inc.

Wock, J.W., & Dirkes, C. (1973). Characteristics of current doctoral dissertations in education. Educational Researcher, 2, 20-22.

Willson, V.L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9 (6), 5-10.

Table 1

Data Set #1

ID	X	Y	X*Y
1	5	1	5
2	10	2	20
3	15	3	45
4	20	4	80
5	25	3	75
6	30	2	60
7	35	1	35
			0
Sum	140	16	320
Count	7	7	8
Mean	20.0000	2.2857	40.0000
Std. Dev.	10.8012	1.1127	30.4725

Table 2

Data Set #2

ID	X	Y	X*Y
1	-1	-1	1
2	-2	-2	4
3	-3	-3	9
4	0	0	0
5	1	1	1
6	2	2	4
7	3	3	9
			0
Sum	0	0	28
Count	7	7	8
Mean	0.0000	0.0000	3.5000
Std. Dev.	2.1602	2.1602	3.7417

Correlation	1.0000	Y-Intercept	0

Perfect Positive Correlation

Table 3

Data Set #3

ID	X	Y	X*Y
1	1	5	5
2	2	4	8
3	3	3	9
4	4	2	8
5	5	1	5
6			0
7			0
			0
Sum	15	15	35
Count	5	5	8
Mean	3.0000	3.0000	4.3750
Std. Dev.	1.5811	1.5811	3.8891

Correlation	1.0000	Y-Intercept	0

Perfect Negative Correlation

Table 4

Data Set #4

ID	X	Y	X*Y
1	100	36	3600
2	110	42	4620
3	100	26	2600
4	130	22	2860
5	120	54	6480
6	115	44	5060
7	130	29	3770
			0
Sum	805	253	28990
Count	7	7	8
Mean	115.0000	36.1429	3623.7500
Std. Dev.	12.5831	11.3200	1927.1885

Table 5

Data Set #5

ID	X	Y	X*Y
1	10	10	100
2	12	11	132
3	9	12	108
4	8	10	80
5	-10	-10	100
6	-11	-12	132
7	-9	-11	99
	-8	-10	80
Sum	1	0	831
Count	8	8	8
Mean	0.1250	0.0000	103.8750
Std. Dev.	10.3846	11.5264	19.9745

Correlation	0.9918	Y-Intercept	0

Table 6

Data set #6

ID	X	Y	X*Y
1	1	1	1
2	1	2	2
3	1	3	3
4	1	4	4
5	5	10	50
6	5	11	55
7	5	13	65
8	5	14	70
Sum	24	58	250
Count	8	8	8
Mean	3.0000	7.2500	31.2500
Std. Dev.	2.1381	5.2847	31.3221

Table 7

Data Set #7

ID	X	Y	X*Y
1	1	1	1
2	2	2	4
3	3	4	12
4	4	6	24
5	5	10	50
6	6	18	108
7	7	30	210
8	8	45	360
Sum	36	116	769
Count	8	8	8
Mean	4.5000	14.5000	96.1250
Std. Dev.	2.4495	15.6935	127.9960

Figure 1

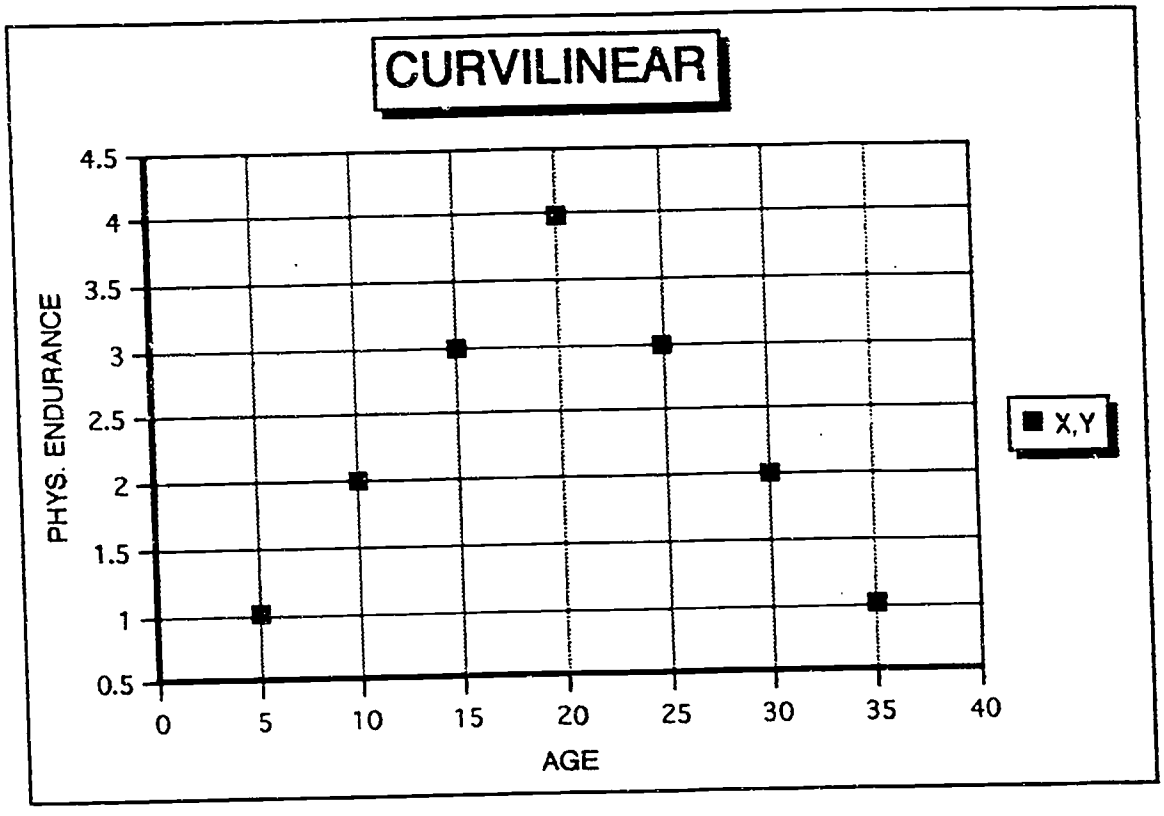

CURVILINEAR
PHYS. ENDURANCE
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
5
10
15
20
25
30
35
40
AGE
X,Y

Figure 2

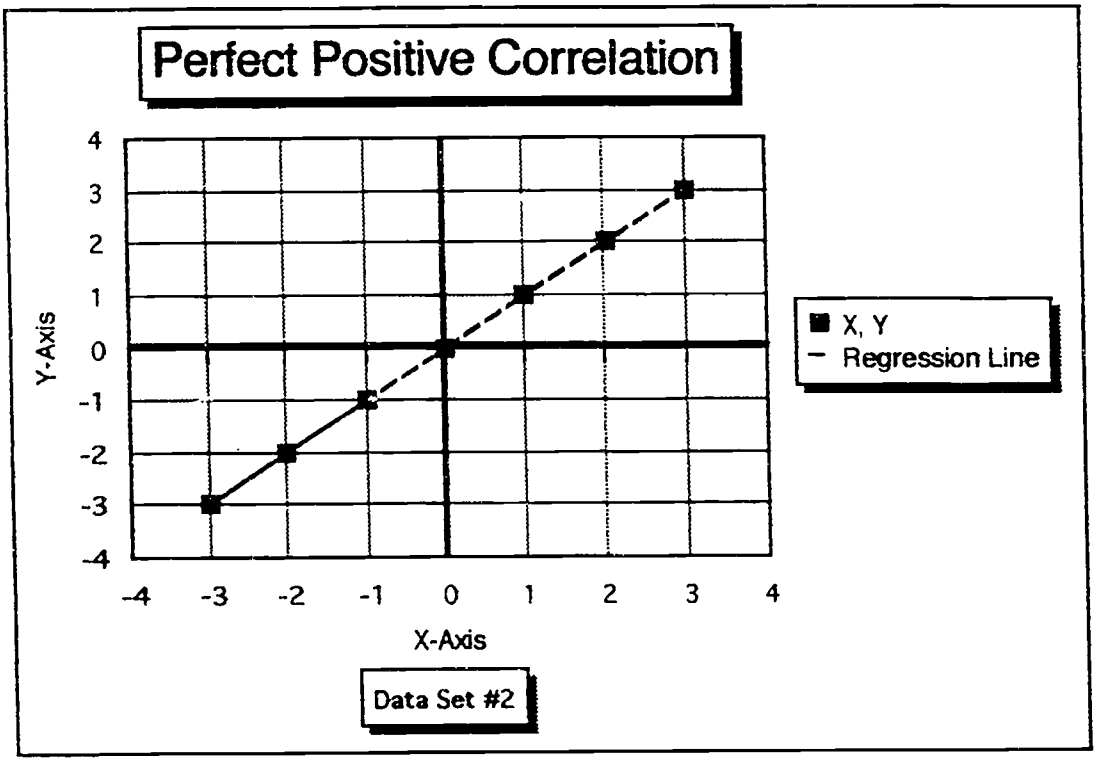

Perfect Positive Correlation
Y-Axis
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
X-Axis
X, Y
Regression Line
Data Set #2

Figure 3

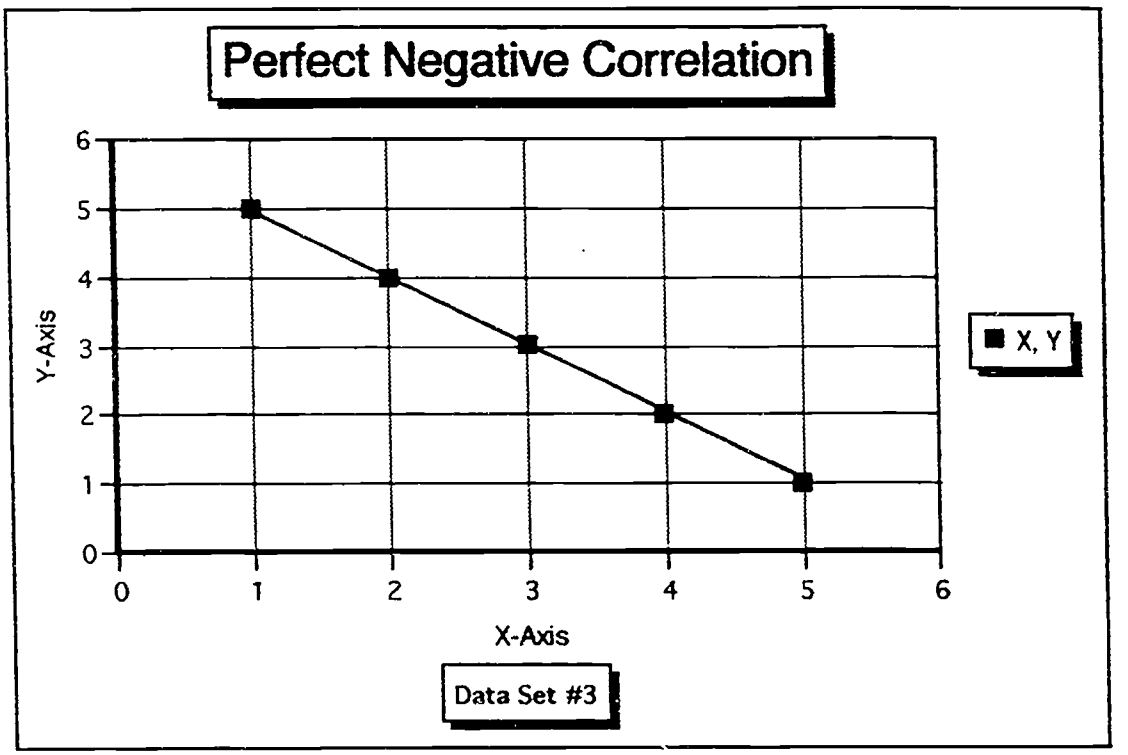

Perfect Negative Correlation
Y-Axis
X-Axis
0
1
2
3
4
5
6
X, Y
Data Set #3

Figure 4

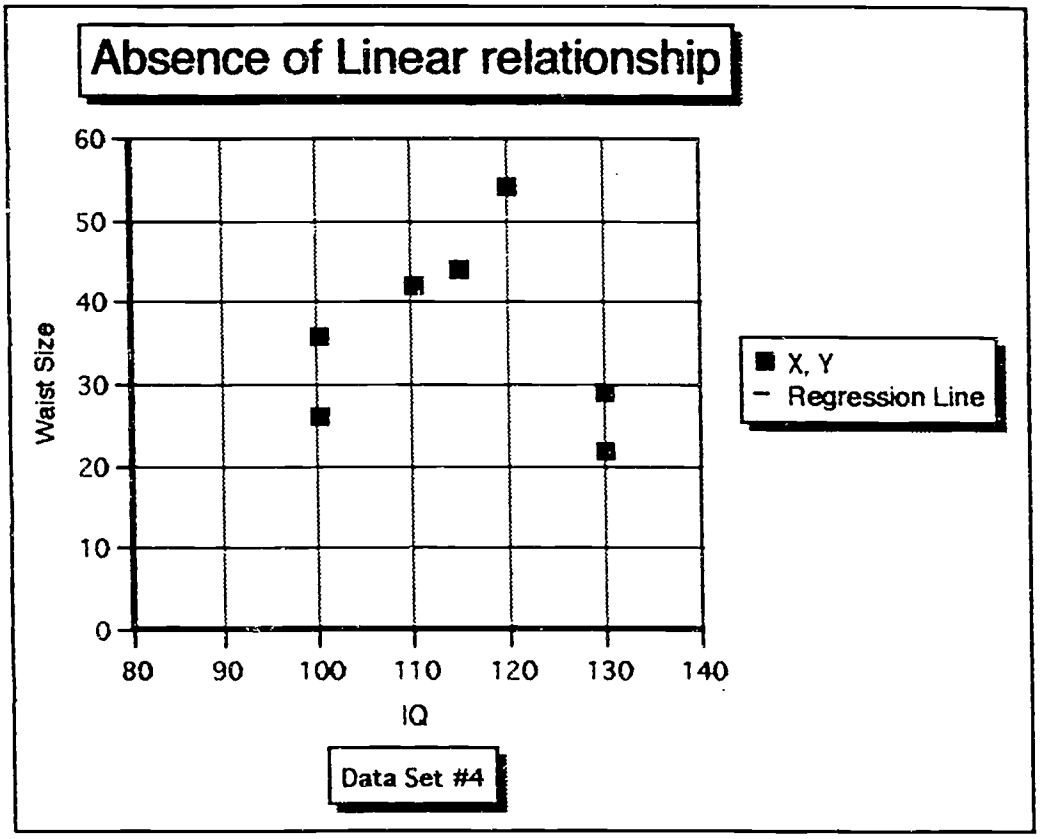

Absence of Linear relationship
60
50
40
30
20
10
0
Waist Size
80
90
100
110
120
130
140
IQ
X, Y
Regression Line
Data Set #4

Figure 5

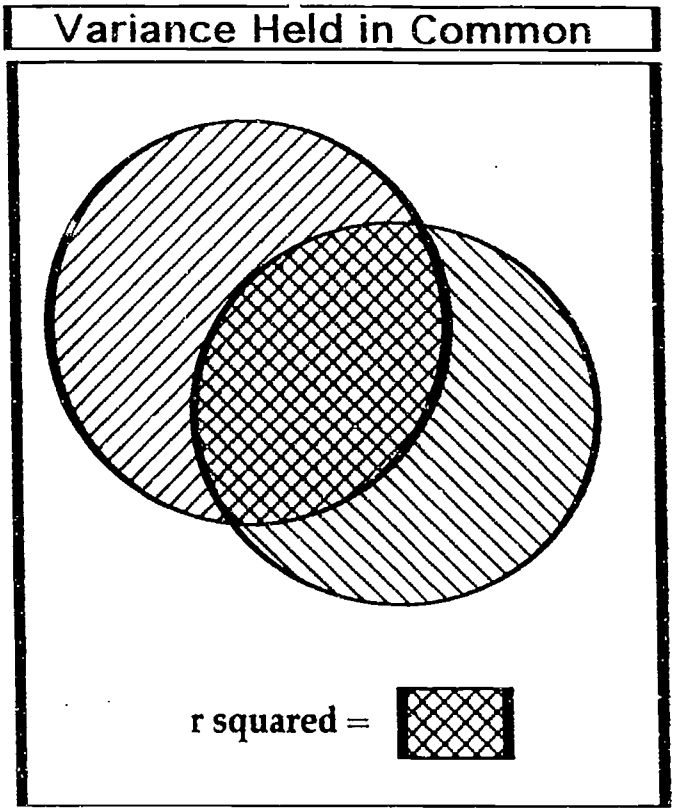
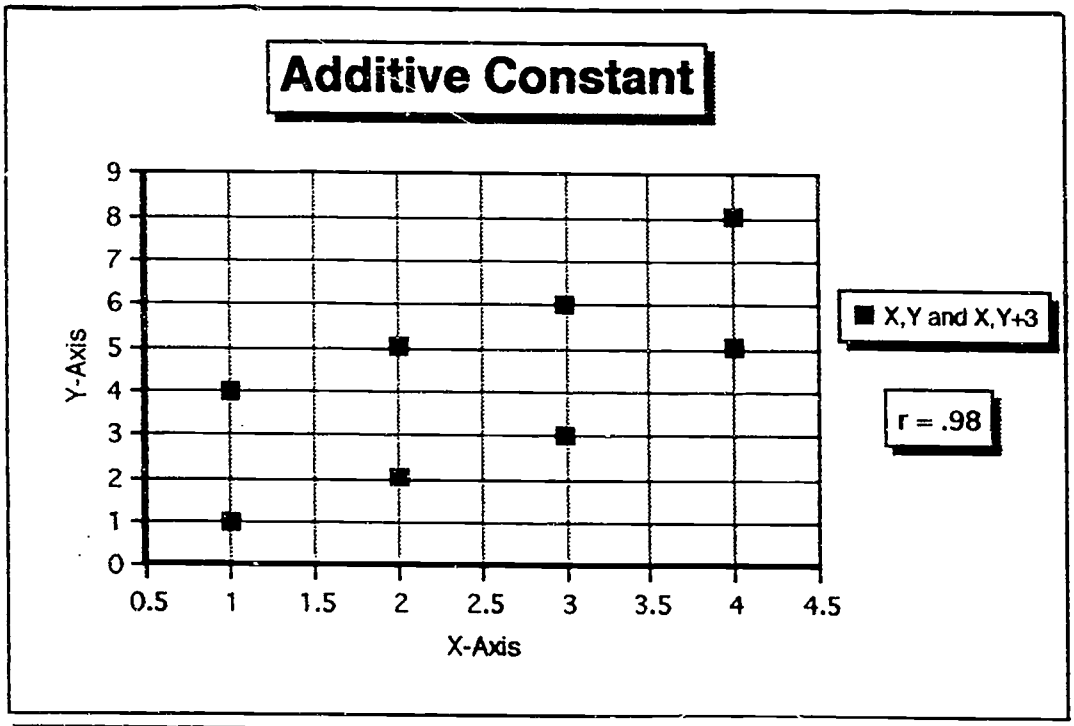


Figure 6


Additive Constant
Y-Axis
X-Axis
X,Y and X,Y+3
r = .98

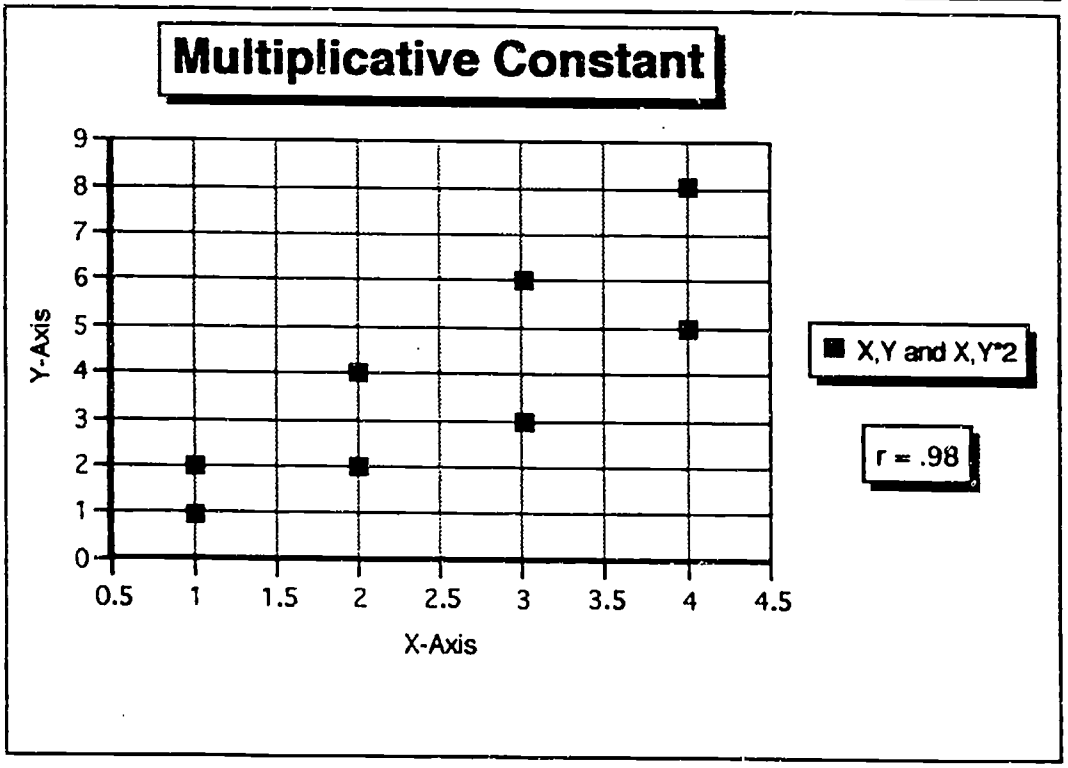
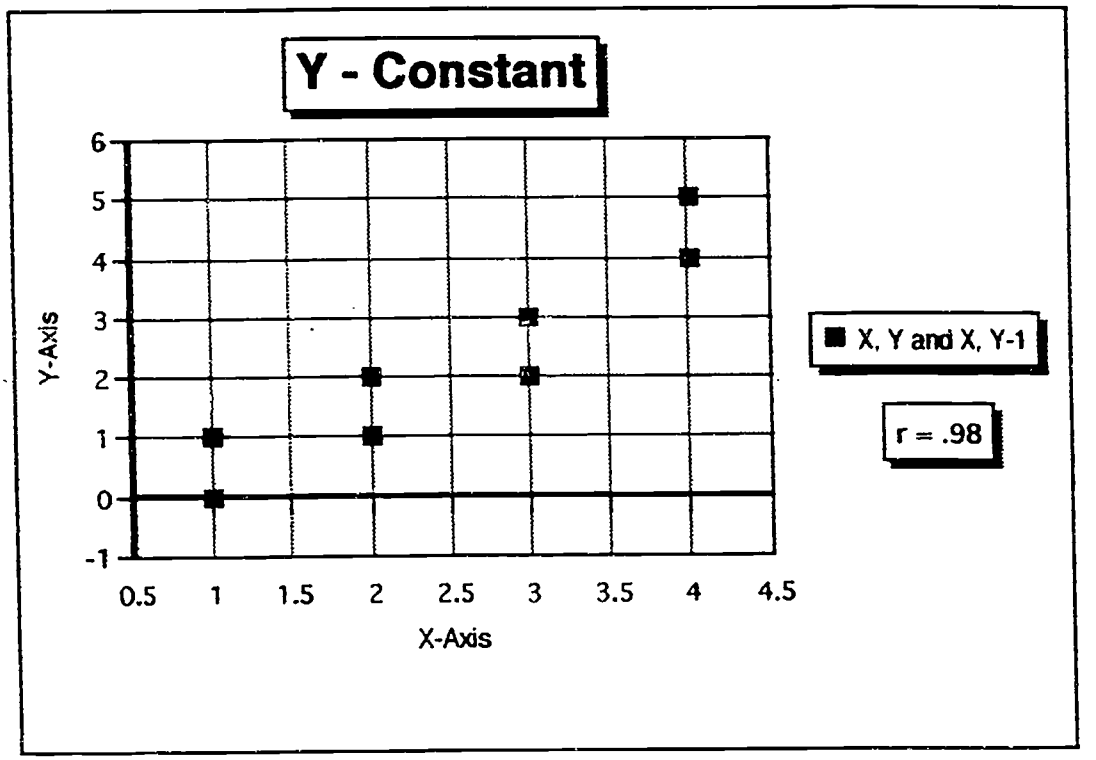

Multiplicative Constant
Y-Axis
X-Axis
X,Y and X,Y*2
r = .98

Figure 6


Y - Constant
6
5
4
3
2
1
0
-1
Y-Axis
0.5
1
1.5
2
2.5
3
3.5
4
4.5
X-Axis
X, Y and X, Y-1
r = .98

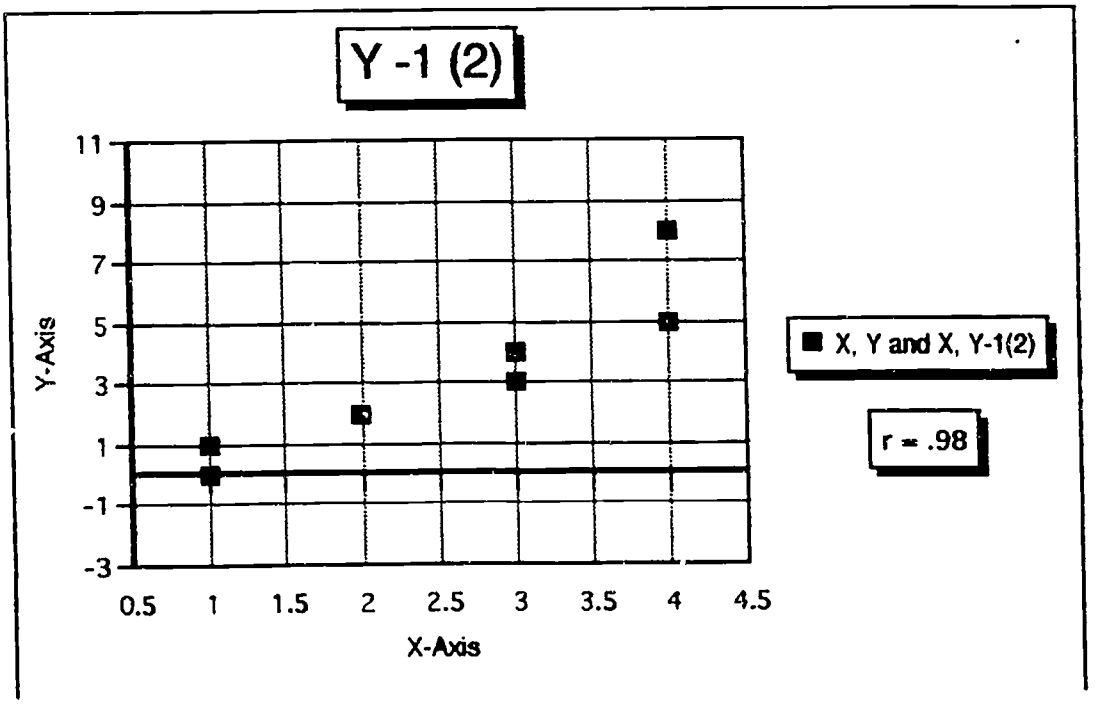

Y -1 (2)
11
9
7
5
3
1
-1
-3
Y-Axis
0.5
1
1.5
2
2.5
3
3.5
4
4.5
X-Axis
X, Y and X, Y-1(2)
r = .98

Figure 7

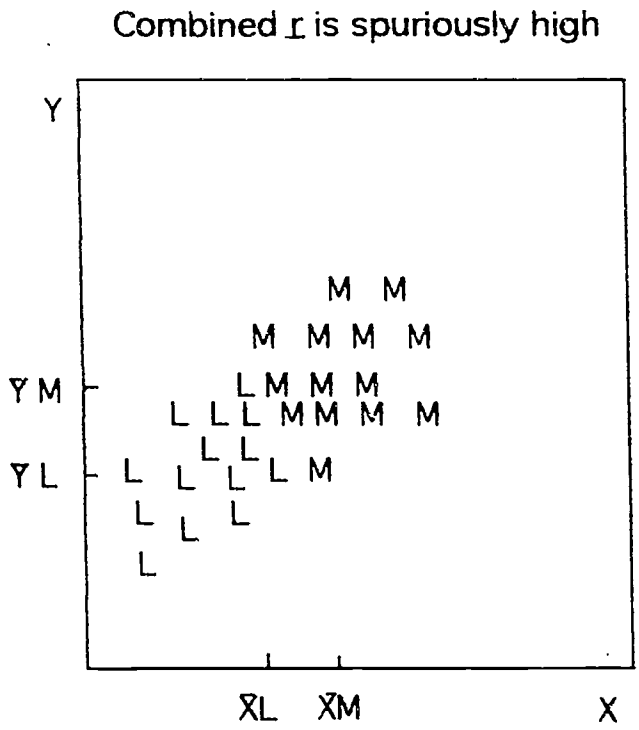
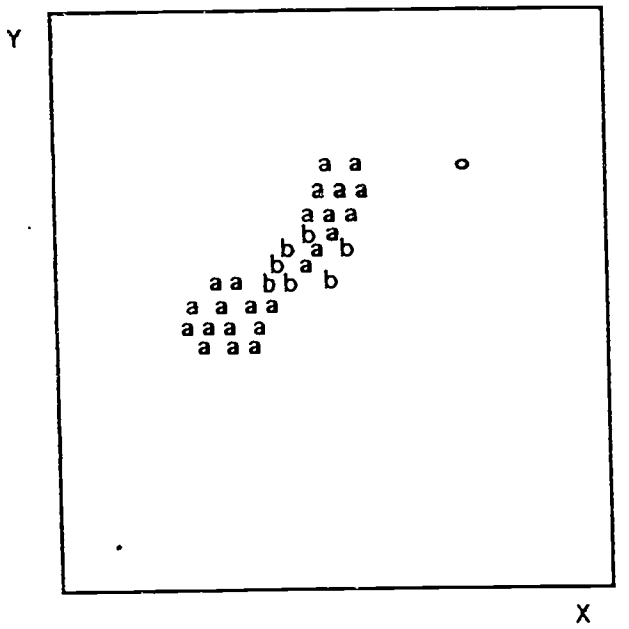

Combined r is spuriously high
Y
Ȳ M
Ȳ L
M M
M M M M
L M M M
L L L M M M M
L L
L L L L M
L L
L
L
X̄L
X̄M
X

Figure 8



Combined r is spuriously high for b and low for a.

Figure 9

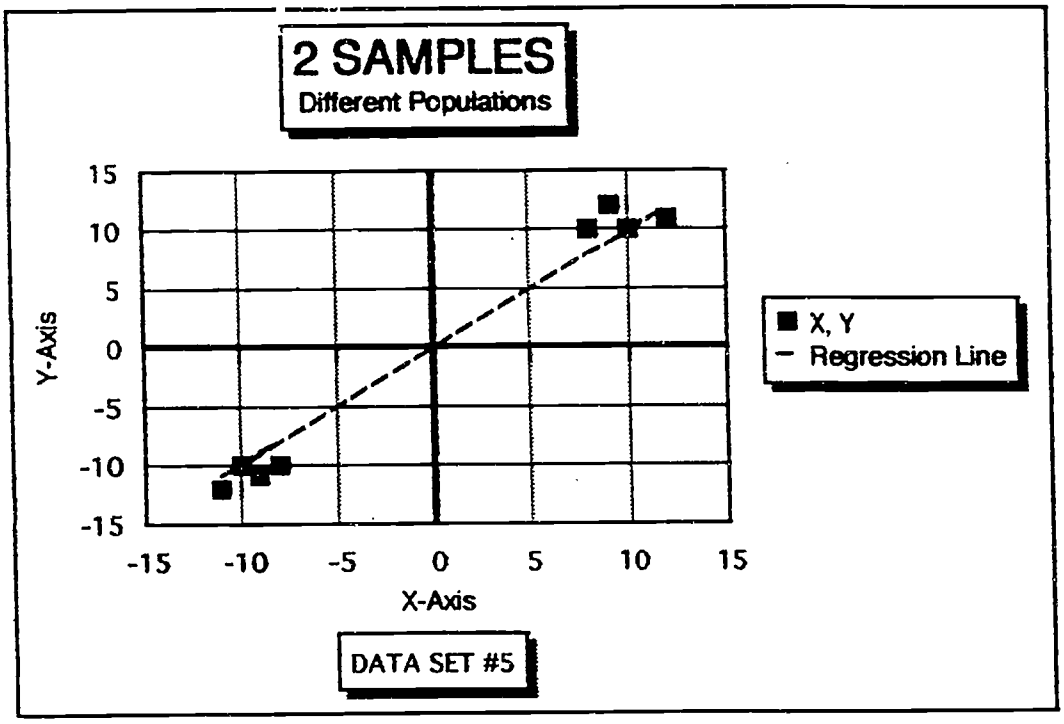
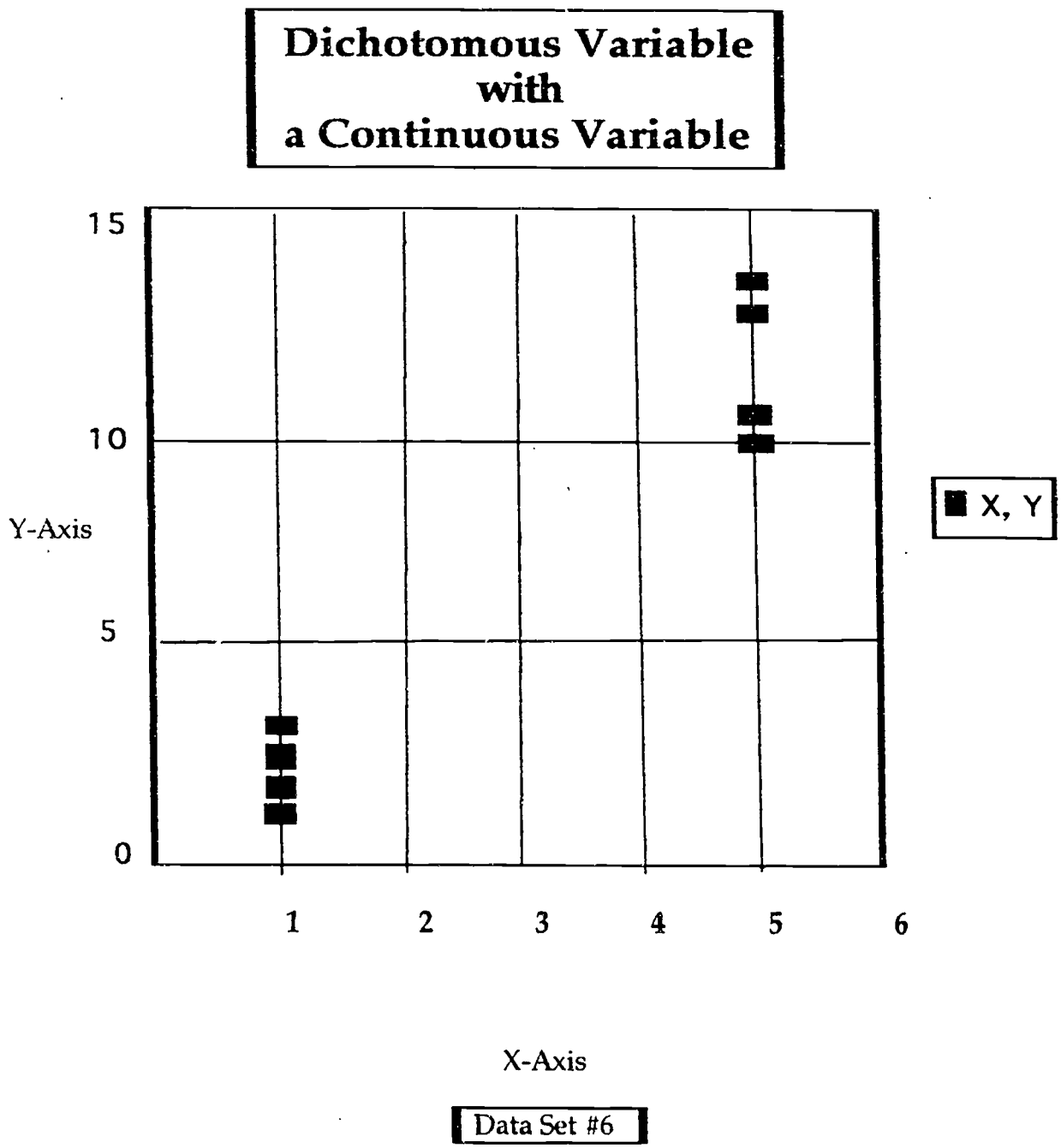
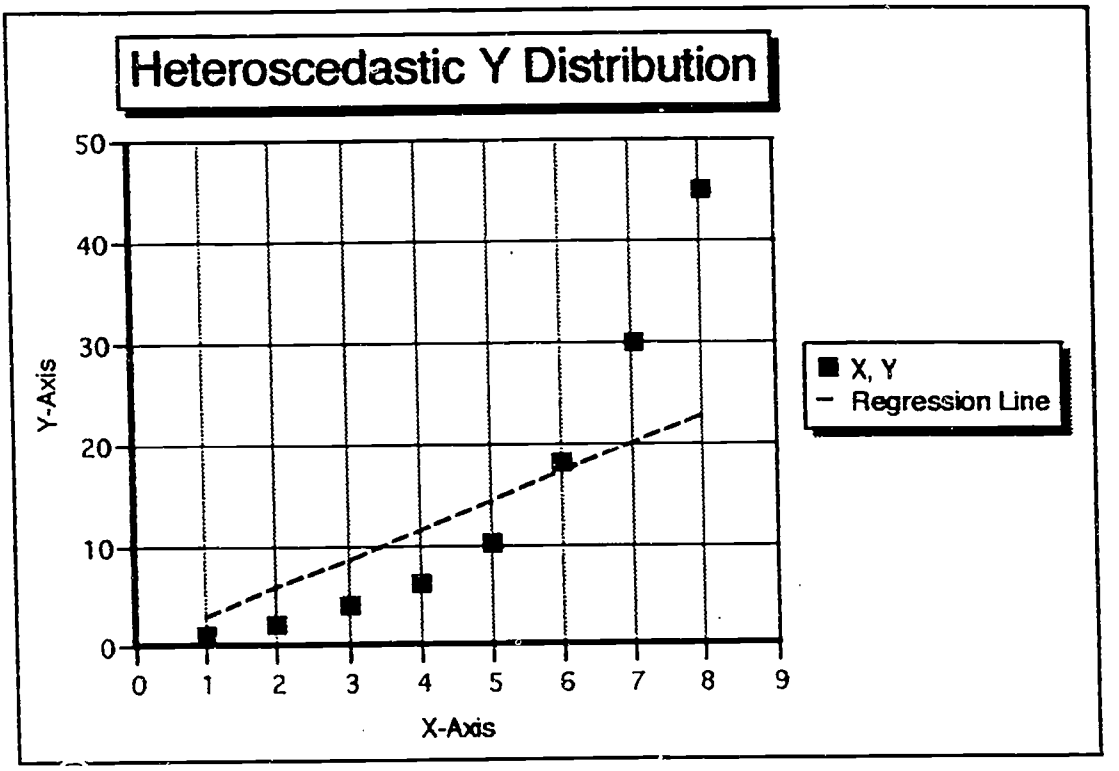

2 SAMPLES
Different Populations
15
10
5
0
-5
-10
-15
Y-Axis
-15
-10
-5
0
5
10
15
X-Axis
X, Y
Regression Line
DATA SET #5

Figure 10



Data Set #6

Figure 11


Heteroscedastic Y Distribution
50
40
30
20
10
0
Y-Axis
0
1
2
3
4
5
6
7
8
9
X-Axis
X, Y
Regression Line